Author’s response to reviews

Title: Voluntary vs. Compulsory Student Evaluation of Clerkships: Effect on Validity and Potential Bias

Authors:

Sola Aoun Bahous (sola.bahous@lau.edu.lb)
Pascale Salameh (pascalesalameh1@hotmail.com)
Angelique Salloum (angelique.salloum@lau.edu.lb)
Wael Salameh (wael.salame@lau.edu.lb)
Yoon Soo Park (yspark2@uic.edu)
Ara Tekian (tekian@uic.edu)

Version: 2 Date: 30 Nov 2017

Author’s response to reviews:

Ms. Ref. No.: MEED-D-17-00327

Title: Voluntary vs. Compulsory Student Evaluation of Clerkships: Effect on Validity and Potential Bias

BMC Medical Education

Dear Dr. Raupach and Distinguished Reviewers:

On behalf of the authors of the manuscript, I would like to thank you for your thoughtful comments and consideration of our manuscript. Please find our responses below. In brief, the authors agreed to make changes based on all suggested comments without reservation.

Sincerely,

Sola Aoun Bahous, MD, PhD, MHPE
Reviewers' Comments

Reviewer 1:

Comment 1:

“Comment 2 Not addressed Authors responded as though this was related to the sham question where intent (which may not have been clear from phrasing) was to query as to why only a half year of compulsory evaluations was compared to a full year of voluntary evaluations. Did this introduce sampling bias?”

Response 1: We thank the reviewer for the clarification. We agree that comparing 2 complete cohorts would have been more meaningful and conclusive. However, the only drawback that we saw in using the whole cohort in the compulsory year is the bias that may be introduced after the sham question was added. We consider that students (especially those who would not have been responders if the evaluation was voluntary) may become more attentive to the content of the subsequent evaluations and become more involved after they were made aware of the sham question; therefore their answers may no longer be independent of the intervention. One of our aims was to examine whether the compulsory approach changes the ratings. While we agree that complete cohorts would be better suited, we want to keep students’ answers unaffected in order to make meaningful comparisons. After administering the sham question, we believe answers to the survey will deviate from pre-administration, which may also introduce a bias. To account for this limitation, we added a note regarding this matter as part of the limitation section, on page 17 (lines 10-14), as presented below:

“Another sampling limitation to be considered is the inclusion of a complete cohort in the first group (voluntary group) and half of a cohort in the second group (compulsory group). This purposeful approach, used to avoid possible effect of the intervention (sham item) on subsequent evaluations, could have introduced a sampling bias.”

Comment 2:

“P4L9 Minor change. "However, few studies..." Need to change to "However, a few studies..." ”

Response 2: This was done with thanks.

Comment 3:

“P9 L17-20 I appreciate the attempt at clarity here but sentence phrasing is still quite awkward and needs refinement. I might suggest the following This intervention was administered to a single cohort during the 2015-16 (compulsory response) academic year. This positioning was
selected to occur after students were familiar with the compulsory process. We felt it was not appropriate to continue with this intervention as students may start to notice the sham question and alter their responses to the entire survey.”

Response 3: We thank the reviewer for the suggestion that clearly makes the reading easier. The suggested sentence was added on page 9, lines 15-19.

Reviewer 2:

Comment 1:

“I had proposed that the authors need to make directional hypotheses, however they suggest in the paper that there is not enough evidence to determine which direction the bias in evaluations was likely to go. I disagree with that somewhat, but I can live with that if their reading of the literature suggests this. However, the Intro section ends in the most tentative way in this draft with the statement, "We suggest that inducing responses using authority does not provide accurate information about the clerkship;" and then in the Methods section (beginning with p8/line 21) the authors begin offering hypotheses. Such hypothesizing very definitely belongs at the end of the Intro section (rather than that general statement about what the authors suggest.”

Response 1: We thank the reviewer for the clarification and we agree that we have suggested a hypothesis related to the effect of a compulsory approach on validity. This hypothesis is now included in the introduction, replacing the previous general statement (page 5, lines 17-19). The introduction now ends with the following sentence:

“We hypothesize that increasing responses using a compulsory approach to evaluation would introduce a quality bias when enforced students (who would elect not to participate) provide inattentive responses.”

Comment 2:

“While they might have stated this directly, the authors made me do the math to determine that the responses came from 49 students who each were asked to rate 7 clerkships. First, the authors never tell us, when responses were voluntary, the extent to which there was consistency across clerkships by individual as to response/nonresponse (or consistency in response to the sham question). For instance, did some individuals rarely or never respond to any of the clerkships, or were different people accounting for nonresponse according to the clerkship. The answer to this question will surely determine how we interpret the data. Second, the authors treat the data in the two conditions, voluntary and required, as if the responses were all independent, but actually the responses were nested within individual. This requires a different kind of analysis than that
reported, one that accounts for the non-independence of the responses. The authors must consult a statistical expert/consultant to re-do their analyses so that the tests account for this and are therefore appropriate for their data (in which a small number of individuals account for a large number of responses).”

Response 2: We very much thank the reviewer for this insightful comment and observation (which we missed to clarify in our earlier revision). Indeed, we consulted our statisticians who reviewed again all the data. We will answer this in consecutive points:

1. We agree with the reviewer’s observation that in the voluntary cohort, respondents and non-respondents in one clerkship may be different from those in another clerkship, and that reporting consistency of response status is very interesting. This analysis was done and yielded a Cronbach’s Alpha of 0.84, which is consistent with the fact that respondents tend to be consistent across clerkships, and vice versa. This was added to the statistics (page 10, lines 22-23), results (under “Response rates and characteristics of responses in the voluntary group”, page 11, lines 16-17) and discussion (page 14, lines 20-22) sections. The additions are presented below:

Under Statistical Analysis: “Cronbach’s alpha was used to determine consistency of response/non-response per individual across clerkships in the voluntary group, and as a reliability estimate.”

Under Results section - Response rates and characteristics of responses in the voluntary group (Table 1): “The consistency of respondent/non-respondent status was 0.84.”

Under Discussion: “Furthermore, there was a consistency of participation per individual across clerkships, whereby responses were largely provided by the same students in each clerkship.”

2. The sham question was administered to the compulsory group (which is a consecutive cohort, and not the same cohort of students). All students answered the sham question and their responses were analyzed and confirmed in a second survey, as described on page 9, lines 19-22, and page 10, lines 1-2. Since the sham question was administered only once, consistency of responses was not considered in this case.

3. To link different responses of the same student, we need to identify the individuals. However, one of the issues involved in this problem is that our respondent identities are anonymous, and as such, a more direct analysis using for example, mixed-effects regression, which accounts for dependence of responses could not be easily conducted. The reviewer’s
comment is valid, and we agree with this issue. Nonetheless, this type of analysis applies to within-cohort comparisons (involving dependent samples), and to between-cohorts comparisons involving large samples obtained from repeated responses. Therefore, comparisons of clerkships involving the voluntary group are concerned, in addition to comparisons involving presence/absence of comments, where the total number of responses was used. Other comparisons involved independent samples. As stated above, our data was not collected in such a manner to allow to identification of the subject (evaluations are anonymous). The only way to do so would have been to lose the anonymity of the data collection process (or devise a system that would a priori plan for it). As of now, we are not able to identify the individual responses per clerkship and thus are forced to analyze the data as non-repeated. The impact of this analysis would mean that the measures of the sum of squares residuals are over estimated (since they will include the residuals and the within), that would lead to a bias F statistics. However, to our knowledge, the bias is towards the null (lower power), and thus we may be able to still conclude that, when our analysis was able to detect statistical significance, the corrected analysis (taking into account the repeat), would declare significance as well, as long as the analysis does not involve the larger sample obtained from repeated responses. Concerning comparison of comment availability (using large sample from repeated responses), there was no statistical significance, which makes it unlikely to find a difference in a smaller sample. This clarification was added to the tables, and to the statistical analysis (page 10, lines 12-21) and discussion (page 17, lines 17-21) sections as shown below:

Statistical Analysis: “However, since responses in different clerkships within the same cohort are dependent, instead of being independent (because common students rate the different clerkships), a more direct analysis using for example mixed-effects regression should have been used. This statistical analysis necessitates the identification of individual responses per student as repeated measures, which is impossible in our study design, where participation in clerkship evaluations is anonymous. Therefore, we analyzed our data as non-repeated, knowing that this reduces the power to detect a difference but does not normally lead to type I error as long as the analysis does not involve the larger sample obtained from repeated responses.“

Discussion: “A fourth and major limitation is related to the statistical analysis involving clerkship comparisons within the same cohort. Conducted analyses, enforced by the study design, could not account for non-independency of provided ratings, hence reducing the power to detect a difference between clerkships. However, this aspect does not involve the key findings of our study.”

Comment 3:
“Finally, the authors are terribly inconsistent in reporting whether there were differences between compulsory vs required in terms of whether there were mean differences between the two. First, we are told that "Average ratings were comparable between the voluntary group, unbiased compulsory group, and biased compulsory group. However, ratings were consistently lower in the biased group though this difference did not reach statistical significance." If the means were not statistically significant, then there were no differences, although it is legitimate to say, parenthetically, that there was a "non-significant tendency" for the findings to be lower. Then we are told that "improving response rate using the compulsory approach... added a quality bias to clerkship evaluations." If there were no significant differences, then what is the evidence for bias?? Finally, in their conclusion, the authors say, "we propose that using authority to improve response rates... could threaten the validity by affecting the quality of rating." I am just not sure what the authors mean here? The writing around the key finding of this paper needs considerable cleaning up so that the message of the findings is both clear and consistent.”

Response 3: We agree with the reviewer that the final message could be confusing throughout the manuscript. In fact, we acknowledge that differences of ratings between voluntary and compulsory modalities were not statistically significant; however, identifying around third of the class being inattentive to evaluation content (or biased), questions the utility of mandating evaluations, and raises the need to confirm this hypothesis in larger, more powerful studies, that can reveal significance, if any. Until then, we agree that the message cannot be strong in favor of bias. We made corrections to make the message clearer and consistent as shown below:

Beginning of discussion (page 13, lines 13-16): “The major findings from this study are as follows: 1) the reliability of ratings was adequate despite the low response rate, and 2) improving response rate using the compulsory approach did not improve reliability, and was associated with inattentive responses in 32.6% of cases without yielding different ratings.”

Discussion (page 15, lines 14-17): “Comparing these evaluations to others provided by non-biased participations and those from the voluntary group, we did not find any statistically significant difference (despite a non-significant tendency toward lower ratings of clerkships in the biased group, and toward a lower availability of comments).”

Discussion (page 15, lines 22-23 and page 16, lines 1-3): “However, this is the first study to our knowledge that examined the effect of a compulsory approach on clerkship evaluation and to show that around a third of responders acknowledged that they provided random, inattentive ratings. Interestingly, These ratings were not significantly different from unbiased and from voluntary participations.”
Conclusion (page 18, lines 6-11): “While response rates and characteristics of non-respondents should be examined before data interpretation is conducted, we propose that using authority to improve response rates may not always improve the reliability, does not yield different ratings, and could threaten the validity only if enforced evaluations were significantly different from voluntary evaluations. The latter needs to be confirmed in larger studies.”